Variable selection in near infrared spectra for the biological characterization of soil and earthworm casts

Lauric Cécillon[1,2,*], Nathalie Cassagne[3], Sonia Czarnes[4], Raphaël Gros[5], Jean-Jacques Brun[1]

Adresses:

[1] Cemagref Grenoble, Mountain Ecosystems Research Unit, 2 rue de la Papeterie, BP 76, 38402 Saint Martin d'Hères, France

[2] French Agency for Environment and Energy Management (ADEME), 2 square La Fayette, BP 90406, 49004 Angers cedex 01, France

[3] INRA, URFM, Ecologie des Forêts Méditerranéennes, UR 629, Domaine Saint Paul, Site Agroparc, 84914 Avignon cedex 9, France

[4] Université de Lyon, Université Lyon 1, CNRS UMR5557; INRA USC 1193, Ecologie Microbienne

Bât G. Mendel, 43 bd du 11 novembre 1918, 69622 Villeurbanne, France

[5] Institut Méditerranéen d'Ecologie et de Paléoécologie (IMEP), Facultés des Sciences St-Jérôme, Boîte 452, 13397 Marseille cedex 20, France

*corresponding author: L. Cécillon, Phone: + 33 (0)130 799 564, E-mail: cecillon@cetiom.fr, Webpage: http://lauric.cecillon.free.fr/

Abstract

Near infrared reflectance spectroscopy (NIRS) was used to predict six biological properties of soil and earthworm casts including extracellular soil enzymes, microbial carbon, potential nitrification and denitrification. Partial least squares regression (PLSR) models were developed with a selection of the most important near infrared wavelengths. They reached coefficients of determination ranging from 0.81 to 0.91 and ratios of performance to deviation above 2.3. Variable selection with the variable importance in the projection (VIP) method increased dramatically the prediction performance of all models with an important contribution from the 1750-2500 nm region. We discuss whether selected wavelengths can be attributed to macronutrient availability or to microbial biomass. Wavelength selection in NIR spectra is recommended for improving PLSR models in soil research.

Keywords: NIRS; PLSR; Wavelength selection; Soil biogenic structures; Soil macrofauna; Microbiological activity

Near infrared reflectance spectroscopy (NIRS) is a rapid and non-destructive analytical technique involving diffuse reflectance measurement in the near infrared (NIR) region (1000-2500 nm). NIR spectra depend on the number and type of chemical bonds in the analysed material (Foley et al., 1998). NIRS is now widely used to predict soil carbon and nitrogen content (e.g. Cécillon and Brun, 2007), usually using partial least squares regression (PLSR) applied over the whole spectrum which is not an optimal solution according to various chemometric studies (e.g. Roger and Bellon-Maurel, 2000). Recent developments of NIRS in Soil Ecology have shown its efficiency to predict soil biological properties including microbial biomass, potential nitrification, soil enzyme activities (Reeves et al., 2000; Cohen et al., 2005; Terhoeven-Urselmans et al., 2008) and to discriminate soil biogenic structures (e.g. earthworm casts) from surrounding soil (Hedde et al., 2005; Velasquez et al., 2007). Hitherto, the predictive capacity of NIRS for biological attributes of soil biostructures has not been assessed. In addition, the prediction of soil microbiological activity like soil enzyme activities with NIRS is not well-established and needs more research.

The objectives of this study were to assess (i) the potential of NIRS to predict various biological attributes of soil and earthworm casts; (ii) the efficiency of selecting the most important NIR wavelengths (process variables which really have an effect on the response) for improving PLSR models.

A sample set was collected in the Maures mountains (Var, France) during spring 2007 in 25 plots depicting a heterogeneous mosaic of Mediterranean forests. This mosaic is representative of a large range of Mediterranean forest ecosystems generated by various wildfire frequencies. The sample set comprises as much samples of earthworm casts (sampled at the soil surface) as topsoil samples (0-5 cm) which could contain some below-ground casts. In each plot, one composite sample was made for topsoil and another one for casts, resulting in a total of 49 samples (Table 1). Casts were mainly produced by *Nicodrilus nocturnus*, a deep-burrowing species found in these forest soils classified as Cambisols (IUSS Working Group WRB, 2006).

Organic carbon, total nitrogen content and pH were determined from conventional analyses. Biological analyses included microbial carbon (C_{mic}), two extracellular enzymes (FDA hydrolase, cellulase), potential nitrification and potential denitrification. Microbial-to-organic carbon ratio (C_{mic}:C_{org}) was computed from obtained data. The soil analyses performed and their basic statistics are summarised in Table 1. Dry-sieved (2 mm) samples were ground (0.25 mm) to obtain homogeneous powders for NIRS analysis. Diffuse reflectance measurements (1000-2500 nm) were carried out using a Fourier-transform NIR spectrophotometer Antaris II (Thermo electron). Each spectrum comprised 32 averaged scans of the sample. Data were collected at a 0.5 nm resolution resulting in 6224 absorbance values per spectrum.

Inference from spectra was done using PLSR which tackles the autocorrelation of spectral data (Tenenhaus, 1998) with all available samples (casts and soils), since comparing topsoils and biogenic structures was beyond the scope of this paper. Each PLSR model was tuned in order to minimize the root-mean-square error of cross-validation (RMSECV), to maximize the Q^2 value (cross-validated R^2, which gives the predicting ability of the model) and the ratio of performance-to-deviation (RPD) obtained from a full-model leave-one-out cross-validation (X-Val). When using small data sets (40-120 samples) in quantitative multivariate modelling, X-Val gives the best estimate of the predictive performance of an obtained model (Martens and Dardenne, 1998). Tuning options included removing outliers detected from a principal component analysis (PCA) on NIR data, multiplicative scatter correction (MSC), computing derivatives of spectra and selecting the most important NIR wavelengths using the variable importance on the projection (VIP) method (Tenenhaus, 1998). The VIP method computes a score for each wavelength. VIP scores obtained by the PLSR correspond to an importance measure of each explanatory variable (i.e. wavelength). Since the average of squared VIP scores equals 1 (Chong and Jun, 2005), only influential wavelengths with a VIP score greater than 1 were kept in the model. A new PLSR was then performed with selected wavelengths. Statistical treatments were conducted using R software version 2.6.1 (R Development Core Team, 2007) with ade4 package for PCA (Chessel et al., 2004), signal package for computing derivatives with Savitzky-Golay filters (Fearn, 2000) and pls package for PLSR (Mevik and Wehrens, 2007) with the VIP algorithm of Chong and Jun (2005).

PCA revealed no spectral outliers (data not shown). Some concentration outliers were removed for FDA hydrolase, potential denitrification and C_{mic}:C_{org} (Table 1). For all properties, second derivative of spectra without MSC was chosen as spectral pretreatment. Three PLSR models and their X-Val statistics for each of the six biological properties are presented (Table 2). They correspond to the same spectral processing with zero, one or two successive steps of variable selection with the VIP method.

Predictive performance of PLSR models using the whole spectrum was very poor with RPD close to 1 for most properties (Table 2, Figure 1). One step of variable selection generated reliable models for all properties with RPD above 1.9 (Table 2). Best models were computed with a second step of variable selection (Figure 2); C_{mic} and potential denitrification obtaining "reasonable" (Williams, 1993) X-Val statistics with Q^2 above 0.9 and RPD above 3 (Table 2). These results are in accordance or even better than previously published studies with Q^2 usually ranging from 0.6 to 0.8 for soil biological properties (e.g. Reeves et al. 2000; Rinnan and Rinnan, 2007). They can be considered as positive for such a heterogeneous sampling design, as confirmed by the strong standard deviation of reference data (Table 1).

Implementing the VIP method resulted in the selection of intervals rather than individual wavelengths with an important contribution from the spectral region 1750-2500 nm (Figure 3). For all biological properties, selected intervals contained wavelengths assigned to organic matter (e.g. 1725 nm, 1930 nm) by Ben-Dor et al. (1997) but also most signature wavelengths of microbial biomass identified by Vaidyanathan et al. (1999), except for cellulase (Table 3). Thus, our models for biological properties could reflect not only the changes in macronutrient availability (amount and quality) which control soil biological activity (Palmborg and Nordberg, 1996; Rinnan and Rinnan, 2007) but also information directly related to microbial biomass even if mean C_{mic} represents only 1.5 mg g^{-1} soil (Table 1).

Improvement of models with variable selection is obvious and should be generalized when using PLSR. Despite the increasing use of NIRS in Soil Science, this strategy has been applied only once: Palmborg and Nordberg (1993) selected 15 wavelengths within NIR spectra of soil

samples according to their regression coefficients. But selecting intervals is more robust than working with few selected variables (Höskuldsson, 2001). Our approach of wavelength selection in NIR spectra with the VIP method provided PLSR models with a reliable prediction performance for microbiological properties of soil and earthworm casts. Next step will be the validation of these models on independent samples. One possible application could be mapping soil conditions with fair predictive precision.

Acknowledgements

This study was part of the Forest Focus IRISE project (Coordinator: M. Vennetier, http://irise.mediasfrance.org/). Authors are grateful to J. Grenet, S. Figarol, F. Ruaudel, N. Guillaumaud, Y. Paillet, B. Couchaud for their technical support and to participants of Grenoble workshop "NIRS in Soil Science" (http://spirsolgrenoble2007.free.fr/) for their valuable comments. This work was supported by the French Agency for Environment and Energy Management (ADEME) and Cemagref.

References

Adam, G., Duncan, H., 2001. Development of a sensitive and rapid method for the measurement of total microbial activity using fluorescein diacetate (fda) in a range of soils. *Soil Biology and Biochemistry* **33**, 943-951.

Anderson, J.P.E., Domsch, K.H., 1978. A physiological method for the quantitative measurement of microbial biomass in soils. *Soil Biology and Biochemistry* **10**, 215-221.

Beare, M.H., Neely, C.L., Coleman, D.C., Hargrove, W.L., 1990. A substrate-induced respiration (sir) method for measurement of fungal and bacterial biomass on plant residues. *Soil Biology and Biochemistry* **22**, 585-594.

Ben-Dor, E., Inbar, Y., Chen, Y., 1997. The reflectance spectra of organic matter in the visible near-infrared and short wave infrared region (400–2500 nm) during a controlled decomposition process. *Remote Sensing of Environment* **61**, 1-15.

Cécillon, L., Brun, J.J., 2007. Near-infrared reflectance spectroscopy (NIRS): a practical tool for the assessment of soil carbon and nitrogen budget. In: Jandl, R., Olsson, M. (Eds.), COST Action 639 : *Greenhouse-gas budget of soils under changing climate and land use (BurnOut)*. Federal Research and Training Centre for Forests, Natural Hazards and Landscape, Vienna, pp. 103-110.

Chessel, D., Dufour, A.B., Thioulouse, J., 2004. The ade4 package - I : One-table methods. *R News* **4**, 5-10.

Chong, I.G., Jun, C.H., 2005. Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems* **78**, 103-112.

Cohen, M.J., Prenger, J.P., DeBusk, W.F., 2005. Visible-near infrared spectroscopy for rapid, non-destructive assessment of wetland soil quality. *Journal of Environmental Quality* **34**, 1422-1434.

Deng, S.P., Tabatabai, M.A., 1994. Cellulase activity of soils. *Soil Biology Biochemistry* **26**, 1347-1354.

Fearn, T., 2000. Savitzky-Golay filters. *NIR news* **11**, 6, 14.

Foley, W.J., McIlwee, A., Lawler, I.R., Aragones, L., Woolnough, A., Berding, N., 1998. Ecological applications of near-infrared spectroscopy - a tool for rapid, cost-effective prediction of the

composition of plant and animal tissues and aspects of animal performance. *Oecologia* **116**, 293-305.

Hedde, M., Lavelle, P., Joffre, R., Jiménez, J.J., Decaëns, T., 2005. Specific functional signature in soil macro-invertebrates biostructures. *Functional Ecology* **19**, 785-793.

Höskuldsson, A., 2001. Variable and subset selection in PLS regression. *Chemometrics and Intelligent Laboratory Systems* **55**, 23-38.

IUSS Working Group WRB, 2006. World reference base for soil resources 2006. World Soil Resources Reports No. 103. FAO, Rome, 128 pp.

Lensi, R., Mazurier, S., Gourbiere, F., Josserand, A., 1986. Rapid-determination of the nitrification potential of an acid forest soil and assessment of its variability. *Soil Biology and Biochemistry* **18**, 239-240.

Martens H.A., Dardenne P., 1998. Validation and verification of regression in small data sets. *Chemometrics and Intelligent Laboratory Systems* **44**, 99-121.

Mevik, B.H., Wehrens, R., 2007. The pls package: principal component and partial least squares regression in R. *Journal of Statistical Software* **18**, 2, 1-24.

NF ISO 10390, 2005. Soil quality, determination of pH. AFNOR, 7 pp.

NF ISO 10694, 1995. Qualité du sol, dosage du carbone organique et du carbone total après combustion sèche (analyse élémentaire). AFNOR, 7 pp.

NF ISO 13878, 1998. Soil quality, determination of total nitrogen content by dry combustion ("elemental analysis"). AFNOR, 5 pp.

Palmborg, C., Nordgren, A., 1993. Modelling microbial activity and biomass in forest soil with substrate quality measured using near infrared reflectance spectroscopy. *Soil Biology and Biochemistry* **25**, 1713-1718.

Palmborg, C., Nordgren, A., 1996. Partitioning the variation of microbial measurements in forest soils into heavy metal and substrate quality dependent parts by use of near infrared spectroscopy and multivariate statistics. *Soil Biology and Biochemistry* **28**, 711-720.

R Development Core Team, 2007. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Reeves, III, J.B., McCarty, G.W., Meisinger, J.J., 2000. Near infrared reflectance spectroscopy for the determination of biological activity in agricultural soils. *Journal of Near Infrared Spectroscopy* **8**, 161-170.

Rinnan, R., Rinnan, A., 2007. Application of near infrared reflectance (NIR) and fluorescence spectroscopy to analysis of microbiological and chemical properties of arctic soil. *Soil Biology and Biochemistry* **39**, 1664-1673.

Roger, J.M., Bellon-Maurel, V., 2000. Using genetic algorithms to select wavelengths in near-infrared spectra : application to sugar content prediction in cherries. *Applied Spectroscopy* **54**, 1313-1320.

Smith, M. S., Tiedje, J. M., 1979. Phases of denitrification following oxygen depletion in soil. *Soil Biology and Biochemistry* **11**, 261-267.

Tenenhaus, M., 1998. *La régression PLS*. Editions Technip, Paris, 254 pp.

Terhoeven-Urselmans, T., Schmidt, H., Joergensen, R.G., Ludwig, B., 2008. Usefulness of near-infrared spectroscopy to determine biological and chemical soil properties: Importance of sample pre-treatment. *Soil Biology and Biochemistry* **40**, 1178-1188.

Tiedje, J. M., Simkins, S., Groffman, P. M., 1989. Perspectives on measurement of denitrification in the field including recommended protocols for acetylene based methods. *Plant and Soil* **115**, 261-284.

Vaidyanathan, S., Macaloney, G., McNeil, B., 1999. Fundamental investigations on the near-infrared spectra of microbial biomass as applicable to bioprocess monitoring. *The Analyst* **124**, 157-162.

Velasquez, E., Pelosi, C., Brunet, D., Grimaldi, M., Martins, M., Rendeiro, A.C., Barrios, E., Lavelle, P., 2007. This ped is my ped: Visual separation and near infrared spectra allow determination of the origins of soil macroaggregates. *Pedobiologia* **51**, 75-87.

Williams, P.C., 1993. What does the raw material have to say? *NIR news* **4**, 5, 13.

Yoshinari, T., Hynes, R., Knowles, R., 1977. Acetylene inhibition of nitrous oxide reduction and measurement of denitrification and nitrogen fixation in soil. *Soil Biology and Biochemistry* **9**, 177-183.

Table 1: Laboratory methods of soil analyses and summary statistics of reference data

Property	n*	Mean	SD	Reference method
Organic carbon (g kg^{-1})	49	52.2	28.0	NF ISO 10694 (1995)
Total nitrogen (g kg^{-1})	49	2.94	1.43	NF ISO 13878 (1998)
PH in H_2O	49	6.42	0.34	NF ISO 10390 (2005)
Potential denitrification (µg N g^{-1} dw h^{-1})	48 (1)	0.24	0.20	Yoshinari et al.(1977); Smith and Tiedje (1979); Tiedje et al. (1989)
Potential nitrification (µg N g^{-1} dw h^{-1})	36**	0.19	0.27	Lensi et al. (1986)
Microbial carbon (mg g^{-1})	49	1.5	0.65	Anderson and Domsch (1978); Beare et al. (1990)
C_{mic}:C_{org} ratio	47 (2)	0.031	0.006	
Cellulase (U_1 g^{-1} dw)	49	0.015	0.007	Deng and Tabatabai (1994)
FDA hydrolase (U_2 g^{-1} dw)	48 (1)	0.0012	0.0003	Adam and Duncan (2001)

Abbreviations:

n = number of samples used to calculate summary statistics and to perform PLSR

SD = standard deviation; FDA = fluorescein di-acetate; dw = dry weight equivalent;

U_1 = µmol of glucose released per minute; U_2 = µmol of fluorescein released per minute

*: the number of concentration outliers removed is in parentheses. In one of the 25 plots, no earthworm casts could be collected above-ground, hence a total of 49 samples

**: only 36 measurements were available for potential nitrification

Table 2: Prediction of microbiological properties with NIRS : X-Val results

Property	Variable selection	Selected wavelengths (%)	h	Q^2	RMSECV	RPD
Potential denitrification	Whole NIR	100	3	0.38	0.16	1.3
	VIP 1	34	4	0.87	0.07	2.8
	VIP 2	16	4	0.91	0.06	3.4
Potential nitrification	Whole NIR	100	1	0.43	0.20	1.3
	VIP 1	26	3	0.70	0.14	1.9
	VIP 2	13	3	0.89	0.09	3.1
Microbial carbon (C_{mic})	Whole NIR	100	2	0.55	0.43	1.5
	VIP 1	28	3	0.83	0.26	2.5
	VIP 2	13	4	0.90	0.20	3.3
C_{mic}:C_{org} ratio	Whole NIR	100	2	0.14	0.0064	1.0
	VIP 1	35	3	0.73	0.0031	1.9
	VIP 2	16	4	0.84	0.0024	2.5
Cellulase	Whole NIR	100	3	0.18	0.0083	0.9
	VIP 1	33	4	0.72	0.0041	1.9
	VIP 2	12	6	0.81	0.0034	2.3
FDA hydrolase (Fdase)	Whole NIR	100	3	0.03	0.00029	1.1
	VIP 1	34	3	0.78	0.00014	2.2
	VIP 2	15	3	0.84	0.00012	2.6

Abbreviations:

h = number of latent variables used to perform PLSR

Whole NIR = PLSR performed over the whole NIR spectrum

VIP 1 = PLSR performed after one step of variable selection with the VIP method

VIP 2 = PLSR performed after two steps of variable selection with the VIP method

Q^2 = coefficient of determination after a full-model leave-one-out cross-validation

RMSECV = root mean squared error of cross-validation

RPD = ratio of performance-to-deviation (calculated as RPD = SD / RMSECV)

Table 3: Selection of signature wavelengths of microbial biomass after two steps of the VIP method (VIP 2)

	Main absorbance peaks of biomass* (nm)							
Property	**1580**	**1690**	**1730**	**1760**	**2060**	**2270**	**2310**	**2350**
Potential denitrification	N	Y	Y	N	Y	Y	Y	Y
Potential nitrification	N	N	Y	Y	Y	Y	Y	N
Microbial carbon (C_{mic})	N	N	Y	N	Y	Y	Y	Y
C_{mic}:C_{org} ratio	N	Y	Y	N	Y	Y	Y	Y
Cellulase	Y	N	N	N	N	N	N	N
FDA Hydrolase (Fdase)	Y	N	Y	N	Y	Y	Y	Y

Abbreviations:

Y = wavelengths corresponding to biomass peak selected

N = wavelengths corresponding to biomass peak deleted

*: determined by Vaidyanathan et al. (1999)

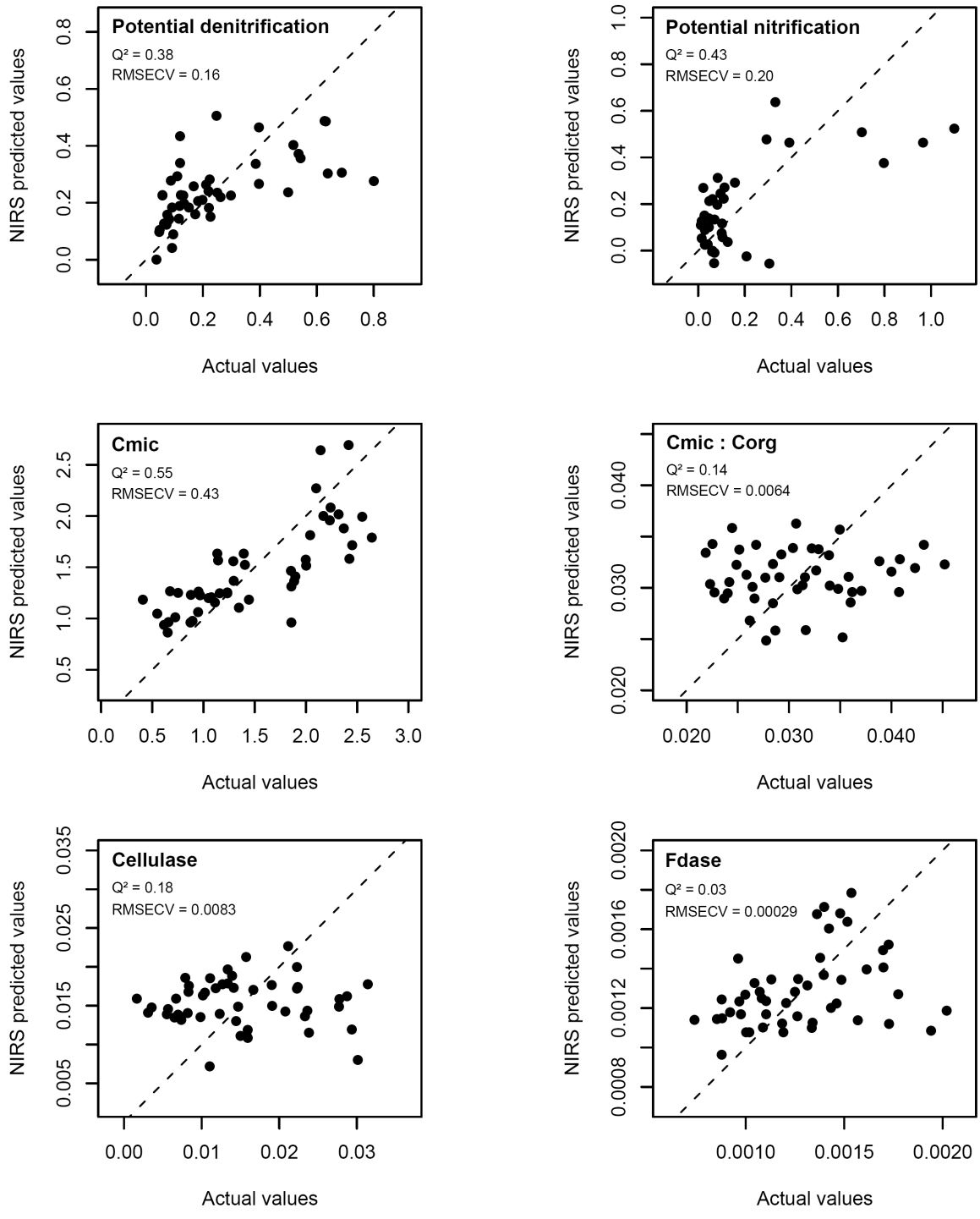

Figure 1: Scatter plots of predicted vs. actual values for biological properties using the whole NIR region. The dashed lines indicate 1:1.

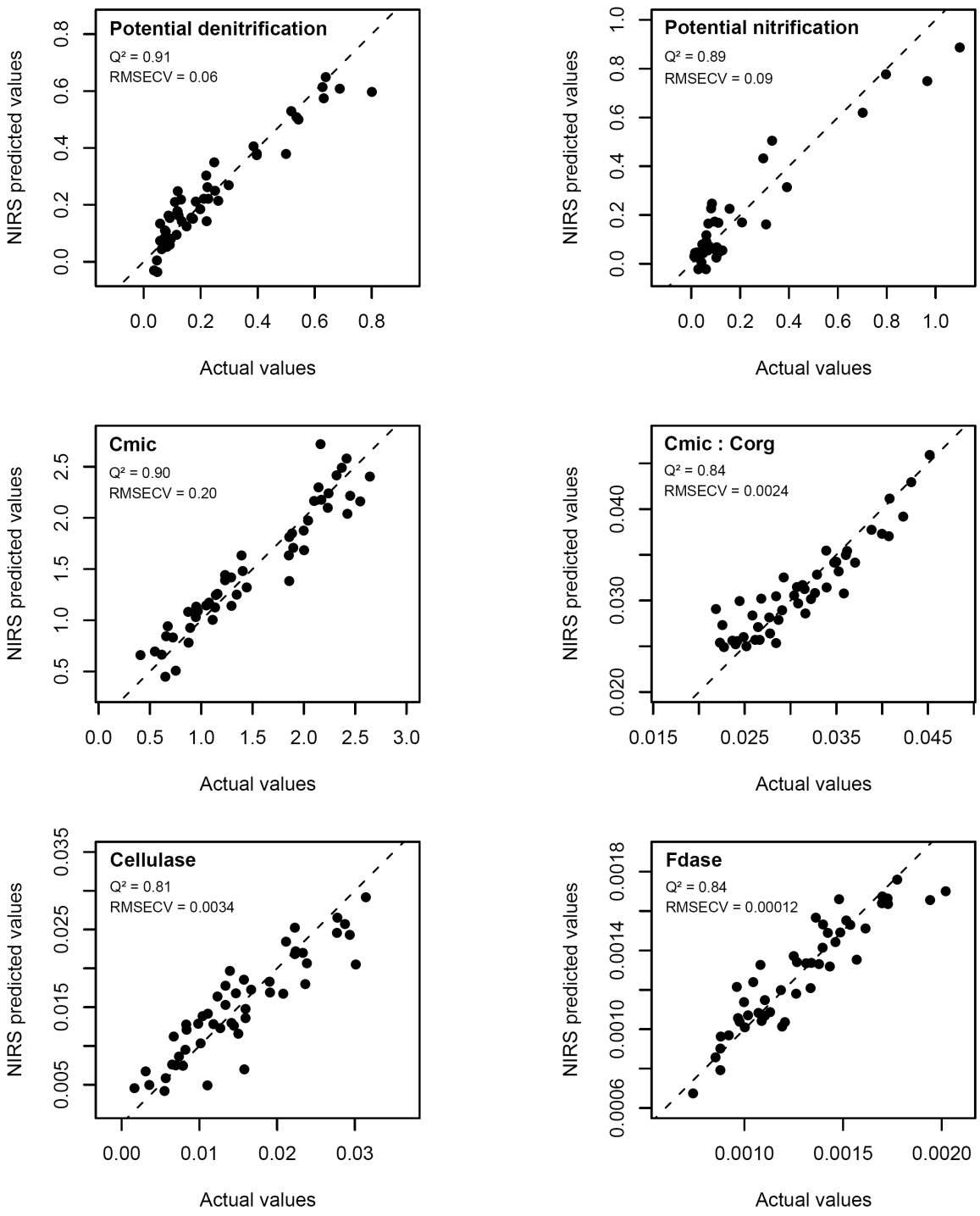

Figure 2: Scatter plots of predicted vs. actual values for biological properties after two steps of variable selection with the VIP method (VIP 2). The dashed lines indicate 1:1.

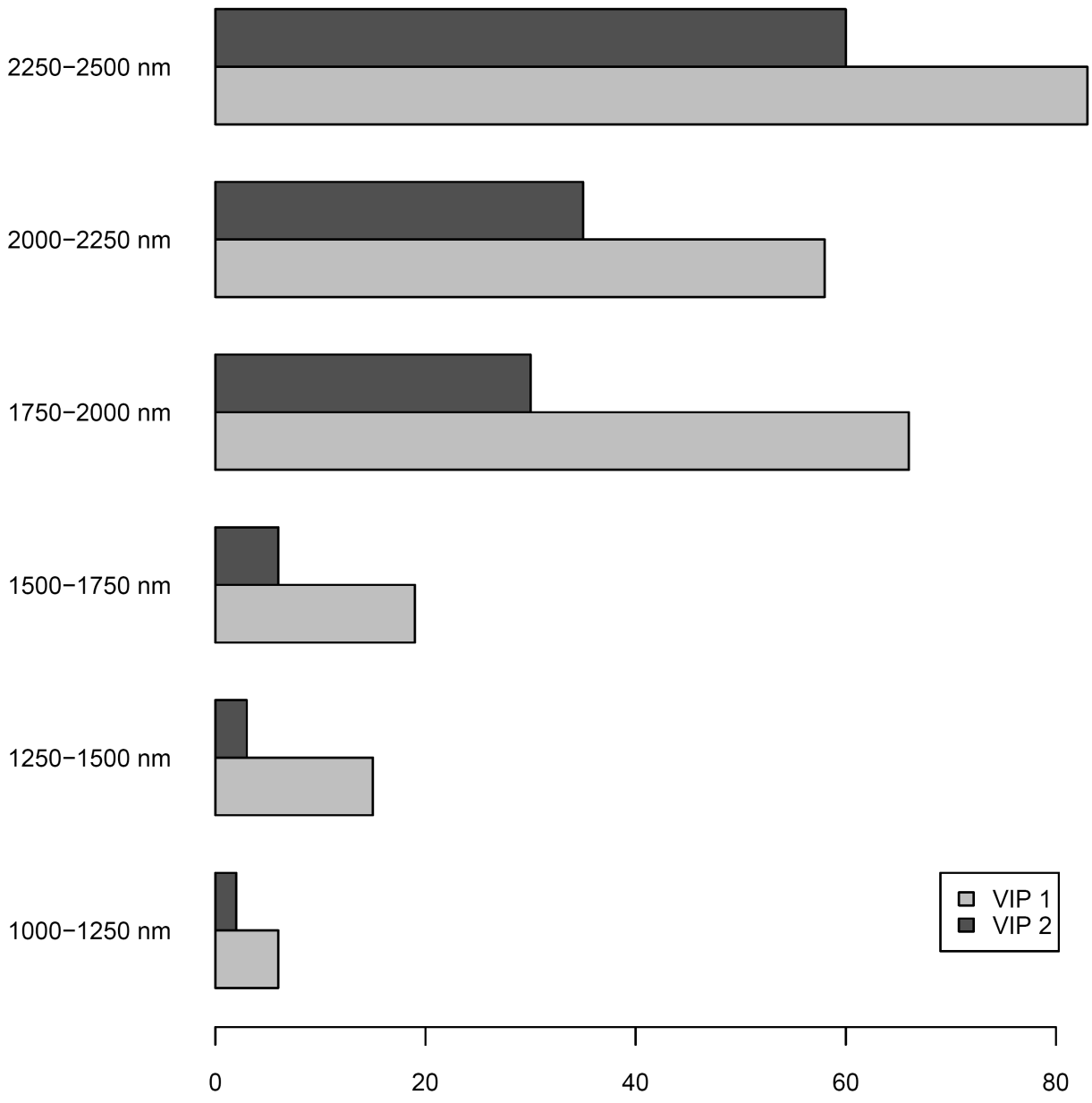

Figure 3: Relative frequency (%) of selected wavelengths within NIR intervals after one (VIP 1) or two (VIP 2) steps of variable selection for microbial carbon* (C_{mic}).

*: similar results were obtained for all properties